0CN

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTIFICATION OF ELECTION

(PCT Rule 61.2)

To:

Commissioner
US Department of Commerce
United States Patent and Trademark
Office, PCT
2011 South Clark Place Room
CP2/5C24
Arlington, VA 22202
ETATS-UNIS D'AMERIQUE

in its capacity as elected Office

Date of mailing (day/month/year) 12 April 2001 (12.04.01)	
International application No. PCT/SG99/00089	**Applicant's or agent's file reference** FP1121
International filing date (day/month/year) 25 August 1999 (25.08.99)	**Priority date** (day/month/year)
Applicant TAN, Ah, Hwee et al	

1. The designated Office is hereby notified of its election made:

[X] in the demand filed with the International Preliminary Examining Authority on:

10 March 2001 (10.03.01)

[ ] in a notice effecting later election filed with the International Bureau on:

2. The election [X] was [ ] was not made before the expiration of 19 months from the priority date or, where Rule 32 applies, within the time limit under Rule 32.2(b).

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland** Facsimile No.: (41-22) 740.14.35	Authorized officer Juan Cruz Telephone No.: (41-22) 338.83.38

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTICE INFORMING THE APPLICANT OF THE COMMUNICATION OF THE INTERNATIONAL APPLICATION TO THE DESIGNATED OFFICES

(PCT Rule 47.1(c), first sentence)

To:
GREENE-KELLY, James, Patrick
Lloyd Wise
Tanjong Pagar
P.O. Box 636
Singapore 910816
SINGAPOUR

LLOYD WISE
13 MAR 2001
RECEIVED

Date of mailing (day/month/year) 01 March 2001 (01.03.01)		
Applicant's or agent's file reference FP1121	**IMPORTANT NOTICE**	
International application No. PCT/SG99/00089	**International filing date** (day/month/year) 25 August 1999 (25.08.99)	**Priority date** (day/month/year)
Applicant KENT RIDGE DIGITAL LABS et al		

1. Notice is hereby given that the International Bureau has communicated, as provided in Article 20, the international application to the following designated Offices on the date indicated above as the date of mailing of this Notice:
   US

   In accordance with Rule 47.1(c), third sentence, those Offices will accept the present Notice as conclusive evidence that the communication of the international application has duly taken place on the date of mailing indicated above and no copy of the international application is required to be furnished by the applicant to the designated Office(s).

2. The following designated Offices have waived the requirement for such a communication at this time:
   EP,SG

   The communication will be made to those Offices only upon their request. Furthermore, those Offices do not require the applicant to furnish a copy of the international application (Rule 49.1(a-bis)).

3. Enclosed with this Notice is a copy of the international application as published by the International Bureau on
   01 March 2001 (01.03.01) under No. WO 01/14992

REMINDER REGARDING CHAPTER II (Article 31(2)(a) and Rule 54.2)

If the applicant wishes to postpone entry into the national phase until 30 months (or later in some Offices) from the priority date, a **demand for international preliminary examination** must be filed with the competent International Preliminary Examining Authority before the expiration of 19 months from the priority date.

It is the applicant's sole responsibility to monitor the 19-month time limit.

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

REMINDER REGARDING ENTRY INTO THE NATIONAL PHASE (Article 22 or 39(1))

If the applicant wishes to proceed with the international application in the **national phase**, he must, within 20 months or 30 months, or later in some Offices, perform the acts referred to therein before each designated or elected Office.

For further important information on the time limits and acts to be performed for entering the national phase, see the Annex to Form PCT/IB/301 (Notification of Receipt of Record Copy) and Volume II of the PCT Applicant's Guide.

The International Bureau of WIPO **34, chemin des Colombettes** **1211 Geneva 20, Switzerland** Facsimile No. (41-22) 740.14.35	Authorized officer J. Zahra Telephone No. (41-22) 338.83.38

ATENT COOPERATION T[illegible]ATY

PCT

REC'D 09 OCT 2001

WIPO PCT

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

(PCT Article 36 and Rule 70)

Applicant's or agent's file reference FP1121	**FOR FURTHER ACTION** See Notification of Transmittal of International Preliminary Examination Report (Form PCT/IPEA/416).	
International application No. PCT/SG 99/00089	International filing date (*day/month/year*) 25 August 1999 (25.08.1999)	Priority Date (*day/month/year*)
International Patent Classification (IPC) or national classification and IPC IPC^7: G06F 15/80; G06K 9/66		
Applicant Kent Ridge Digital Labs et al.		

1. This international preliminary examination report has been prepared by this International Preliminary Examination Authority and is transmitted to the applicant according to Article 36.

2. This REPORT consists of a total of 4 sheets, including this cover sheet.

☐ This report is also accompanied by ANNEXES, i.e., sheets of the description, claims and/or drawings which have been amended and are the basis for this report and/or sheets containing rectifications made before this Authority (see Rule 70.16 and Section 607 of the Administrative Instructions under the PCT).

These annexes consist of a total of ______ sheets.

3. This report contains indications relating to the following items:

I. ☒ Basis of the opinion

II. ☐ Priority

III. ☐ Non-establishment of opinion with regard to novelty, inventive step and industrial applicability

IV. ☐ Lack of unity of invention

V. ☒ Reasoned statement under Rule 66.2(a)(ii) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

VI. ☐ Certain documents cited

VII. ☐ Certain defects in the international application

VIII. ☐ Certain observations on the international application

Date of submission of the demand 10 March 2001 (10.03.2001)	Date of completion of this report 22 June 2001 (22.06.2001)
Name and mailing address of the IPEA/AT Austrian Patent Office Kohlmarkt 8-10 A-1014 Vienna Facsimile No. 1/53424/200	Authorized officer MIHATSEK Telephone No. 1/53424/329

INTERNATIONAL PRELIMINARY EXAMINATION REPORT

International application No.
PCT/SG 99/00089

I. Basis of the report

1. With regard to the **elements** of the international application:*

☒ the international application as originally filed

☐ the description:
pages ______ , as originally filed
pages ______ , filed with the demand
pages ______, filed with the letter of ______.

☐ the claims:
pages ______ , as originally filed
pages ______ , as amended (together with any statement) under Article 19
pages ______ , filed with the demand
pages ______, filed with the letter of ______.

☐ the drawings:
pages ______ , as originally filed
pages ______ , filed with the demand
pages ______ , filed with the letter of ______.

☐ the sequence listing part of the description:
pages ______ , as originally filed
pages ______ , filed with the demand
pages ______ , filed with the letter of ______.

2. With regard to the **language**, all the elements marked above were available or furnished to this Authority in the language in which the international application was filed, unless otherwise indicated under this item.
These elements were available or furnished to this Authority in the following language ______ which is:

☐ the language of a translation furnished for the purposes of international search (under Rule 23.1(b)).

☐ the language of publication of the international application (under Rule 48.3(b)).

☐ the language of the translation furnished for the purposes of international preliminary examination (under Rule 55.2 and/ or 55.3).

3. With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, the international preliminary examination was carried out on the basis of the sequence listing:

☐ contained in the international application in printed form.

☐ filed together with the international application in computer readable form.

☐ furnished subsequently to this Authority in written form.

☐ furnished subsequently to this Authority in computer readable form.

☐ The statement that the subsequently furnished written sequence listing does not go beyond the disclosure in the international application as filed has been furnished.

☐ The statement that the information recorded in computer readable form is identical to the written sequence listing has been furnished.

4. ☐ The amendments have resulted in the cancellation of:

☐ the description, pages ______.

☐ the claims, Nos. ______.

☐ the drawings, sheets/fig ______.

5. ☐ This report has been established as if (some of) the amendments had not been made, since they have been considered to go beyond the disclosure as filed, as indicated in the Supplemental Box (Rule 70.2(c)).**

* *Replacement sheets which have been furnished to the receiving Office in response to an invitation under Article 14 are referred to in this report as „originally filed" and are not annexed to this report since they do not contain amendments (Rules 70.16 and 70.17).*

** *Any replacement sheet containing such amendments must be referred to under item 1 and annexed to this report.*

V. Reasoned statement under Article 35(2) with regard to novelty, inventive step or industrial applicability; citations and explanations supporting such statement

1. Statement

Novelty (N)	Claims	1-17	YES
	Claims		NO
Inventive step (IS)	Claims	1-17	YES
	Claims		NO
Industrial applicability (IA)	Claims	1-17	YES
	Claims		NO

Citations and explanations (Rule 70.7)

The following documents have been cited in the Search Report:

D1: WO 97/04400 A1
D2: US 5566273 A
D3: US 5794236 A
D4: GB 2278705 A
D5: US 5091964 A
D6: JP 11 085796 A
D7: JP 11 039313 A
D8: JP 11 085797 A

Explanations:
The present invention relates to a document classification apparatus comprising feature extraction means for extracting a plurality of features from a document and a classifier operable to process the document in a knowledge acquisition mode in which information for classification is acquired from the features of the document and added incrementally to a knowledge base and in a document classification mode in which the classifier (30), using the knowledge base, determines from the features a predicted classification for the document, the classifier being switchable between the modes under user control. The apparatus is further operable to perform rule insertion in the knowledge acquisition mode in which a number of features are input by a user to the classifier together with a classification with which the features are associated.

<u>WO 97/04400:</u>
The present invention relates to a pattern recognition system that uses an artificial neural network architecture with novel mechanisms for object and 2-D pattern (signal) recognition in cluttered and noisy images. The proposed neural network, hereinafter referred to as Selective Attention Adaptive

Supplemental Box
(To be used when the space in any of the preceding boxes is not sufficient)

Continuation of: Box V (page 1)

US 5566273:
The present invention relates generally to the field of training a neural network and, more particularly, to a system and method that provides a supervised learning environment for a neural network.

US 5794236:
Computer-based system for classifying documents into a hierarchy and linking the classifications to the hierarchy.

GB 2278705:
Facsimile with neural network for character recognition - uses neural networks to recognise characters in scanned or received electronic facsimile image and convert image into characters of standard digital or graphical character set.

US 5091964:
Apparatus for extracting text region in document image - calculates peripheral distribution of filled pixels by projecting pixels in X or Y-axis direction.

JP 11 085796:
Automatic document classification apparatus - has classification decision unit that compares document vector derived for classifying document and folder vector to decide category in which classifying document belongs.

JP 11 039313:
Knowledge base generation method of automatic document classification system - involves eliminating unsuitable characteristic from learning data based on calculated relation degree of characteristic and classification category.

JP 11 085797:
Automatic document-classifying apparatus for e.g. hard disc - has classification decision unit that compares document vector derived for classifying document and folder vector to decide category in which classifying document belongs

Compared with the documents cited in the search report one has to state that although the documents cited show some features of the invention e.g.ART they do not cover all the features of the invention such as knowledge acquisition mode,switchable classifier , ,ARTMAP system,extractions means and so forth.
Furthermore also a combination of the documents cited cannot solve the problem posed.
Therefore the claimed invention seems to be new and inventive.
Industrial applicability is given.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 March 2001 (01.03.2001)

PCT

(10) International Publication Number
WO 01/14992 A1

(51) International Patent Classification[7]: G06F 15/80, G06K 9/66

(21) International Application Number: PCT/SG99/00089

(22) International Filing Date: 25 August 1999 (25.08.1999)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: KENT RIDGE DIGITAL LABS [SG/SG]; 21 Heng Mui Keng Terrace, Singapore 119613 (SG).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: TAN, Ah, Hwee [SG/SG]; Blk 892A, Tampines Avenue 8, #13-30, Singapore 521892 (SG). LAI, Fon, Lin [SG/MY]; 101-G Jervois Road, Singapore 249058 (SG).

(74) Agent: GREENE-KELLY, James, Patrick; Lloyd Wise, Tanjong Pagar, P.O. Box 636, Singapore 910816 (SG).

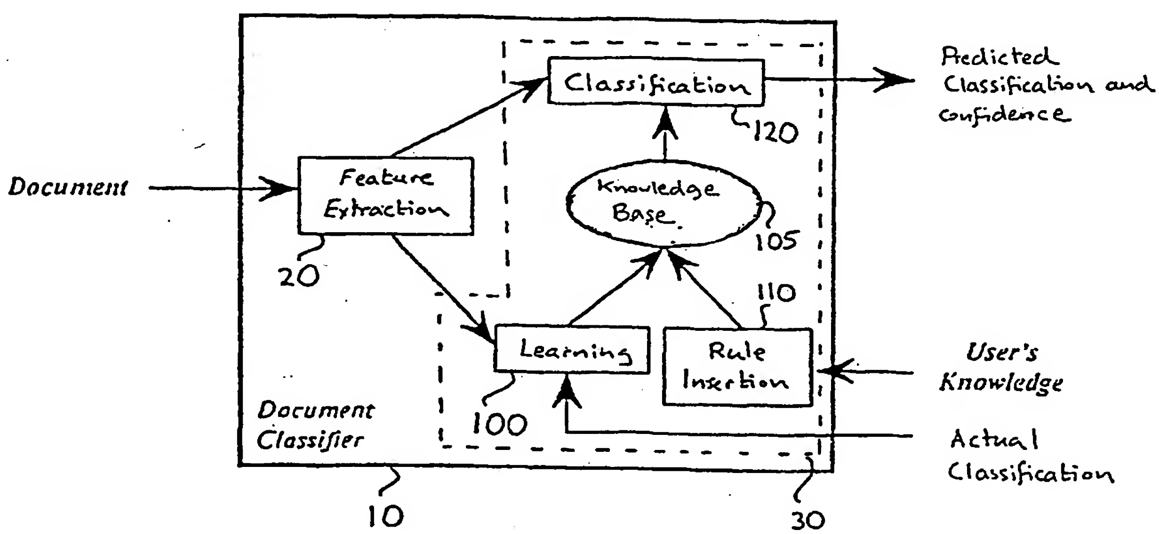
(81) Designated States *(national)*: SG, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published:
— *With international search report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: DOCUMENT CLASSIFICATION APPARATUS



(57) Abstract: Document classification apparatus is disclosed comprising feature extraction means (20) for extracting a plurality of features from a document and a classifier (30) operable to process the document in a knowledge acquisition mode in which information for classification is acquired from the features of the document and added incrementally to a knowledge base (105) and in a document classification mode in which the classifier (30), using the knowledge base, determines from the features a predicted classification for the document, the classifier (20) being switchable between the modes under user control. The apparatus is further operable to perform rule insertion in the knowledge acquisition mode in which a number of features are input by a user to the classifier together with a classification with which the features are associated.

WO 01/14992 A1

DOCUMENT CLASSIFICATION APPARATUS

BACKGROUND AND FIELD OF THE INVENTION

This invention relates to apparatus for classifying documents.

Traditionally, documents which arrive at a central location, for example a post room or facsimile machine and which need to be distributed to a certain destination are sorted and delivered by hand. Efforts have been made to automate this process. For example, it has been proposed in US 5,461,488 to provide apparatus which identifies the destination of a facsimile by applying document image analysis and recognition techniques to the facsimile. US 5,461,488 provides routing based on identification of recipient name. However, for many faxes received, for example in information gathering or public service industries, information identifying a specified recipient may not be present, so that such faxes would not be routed automatically.

General text classifying systems which classify documents into one or more categories have been proposed in US 5,371,807 and US 5,675,710. Such systems use only a single classification strategy, either profile-based, having a keyword/character profile for each category or rule-based in which category knowledge is represented in the form of rules. The systems also use only a single knowledge acquisition strategy, either statistically learned knowledge or user-specified knowledge to provide the knowledge base with which text from a document to be classified is compared to provide the document classification.

It is a disadvantage of the prior art systems noted above that they are prone to mis-classification and consequent mis-routing of documents, as well as cumbersome operation.

It is an object of the invention to provide an improved document classification apparatus.

SUMMARY OF THE INVENTION

According to the invention, there is provided document classification apparatus comprising feature extraction means for extracting a plurality of features from a document and a classifier operable on the extracted features to process the document in a knowledge acquisition mode in which the association of a classification with the document is added incrementally to a knowledge base or in a document classification mode in which the classifier, using the knowledge base, determines a predicted classification for the document, the classifier being switchable between the modes under user control .

The features are preferably formed into a feature vector for input to the classifier and the features preferably comprise classification-associated words or phrases which may appear in the document. The extracting means may be arranged to provide a measure of the frequency of occurrence of the features in the document.

The classifier may comprise a supervised ART system, preferably an ARAM system of the type disclosed in "Adaptive Resonance Associative Map", an article by one of the

present inventors Ah-Hwee Tan, published in "Neural Networks", Vol 8 No 3 pp 437-446. 1995 or an an ARTMAP system of the type disclosed in US 5,214.715.

The apparatus may further be operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

The apparatus may further be operable in a rule insertion sub-mode of the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

The apparatus may further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification and the classification may have associated therewith a confidence value comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison, with a said destination being a system administrator, responsible for manual routing.

The described embodiment provides a document classification apparatus which allows learning to be performed in an incremental way by allowing a system administrator to correct document classification mistakes as they occur, the apparatus learning from these mistakes. By incremental learning of new cases does not require re-learning of previous cases, thus eliminating the need to preserve past cases for re-learning. While the described embodiment focuses primarily on incremental learning, the apparatus is further able to perform learning of a plurality of cases as a batch. During batch learning, the apparatus

learns each case one by one and accumulates the classification information into the knowledge base. Besides learning from training data, the apparatus also allows rules to be inserted into the learning process, leading to a more flexible learning environment.

The apparatus is furthermore able to determine a confidence that the classification of a particular document is correct in the form of a confidence value. This confidence value is compared to a threshold parameter to decide if automatic or manual routing is desirable. Adjustment of this threshold parameter allows the degrees of manual and automatic routing to be controlled, by adjustment of the threshold to match a desired confidence value, thus allowing a smooth transition from a state where manual routing is favoured to one, as the classifier becomes more accurate, that favours automatic routing.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the invention will now be described by way of example with reference to the accompanying drawings in which:

Figure 1 is a schematic diagram illustrating the structure of the described embodiments of the invention;

Figure 2 is a diagram illustrating the document classifier of Figure 1 in a document classification made;

Figure 3 is a diagram illustrating the modes of operation of the embodiments of the invention;

Figure 4 is a diagram of an ARAM system used as a document classifier in an embodiment of the invention.

Figures 5, 6, and 7 summarize the parameter setting and the relevant functional blocks of the document classification system in the learning, rule insertion, and document classification modes respectively.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

With reference to Figures 1-3, a document classification apparatus is shown. The apparatus is operable in a knowledge acquisition mode and a document classification

mode. In knowledge acquisition mode, the apparatus learns from training documents and rules to recognise categories based on document content. This knowledge is then applied in document classification mode to classify further documents. The structure of the apparatus is shown in Figures 1 and 2 and will now be described with reference to the document classification mode. The structure is knowledge acquisition mode in the same, but used differently as described with reference to Figure 3.

A document text file, for example a text file derived from a scanned and OCR processed physical document or derived from a received and stored facsimile message which has been analysed for and converted to textual content, or a word processor document file, is fed to a document classifier 10. The document classifier includes a **feature extraction** module 20 which analyses the text file and extracts previously **selected features** in the form of keywords or phrases from that file which are fed as a feature vector to a **classifier** 30 which is in the form of an ARAM (Adaptive Resonance Associative Map) system which provides a predicted classification for the output document in response to the input feature vector. This classification is associated with a **confidence value** which, together with the document, is passed to a router 40. At the router 40, the value is compared to a threshold input by a system administrator 50. If the value exceeds the threshold, the document is routed to the destination 52 specified by the classification, via path 55. If not, the document is routed to the system administrator 50 for manual routing via path 60. The destinations 52 can also communicate with the system administrator 50 through path 60, to return misdirected documents for manual routing.

The modes of operation of the apparatus are shown in Figure 3. **In knowledge acquisition mode**, two sub-modes are used. The first, represented by block 100, is based on **learning** and requires the input of training data in the form of documents, for each of which a feature vector is extracted by module 20 and fed to module 30. The training documents can either be input individually or as a batch. The actual category of the document is input by system administrator 50 and fed to the module 30. Module 30 then adjusts (if necessary) the association of the vector to the predicted category in a knowledge base 105 so that the predicted category equals the actual category. The second sub-mode is based on **rule insertion**, represented by block 110. In rule insertion, a feature vector and an actual category are input by the system administrator 50 and an association between the input vector and the actual category is made, if one does not already exist.

In **document classification mode**, represented by block 120, the feature vector and document are fed to the module 30 and based on the knowledge acquired by the knowledge base in the knowledge acquisition mode, a classification is determined in accordance with the feature vector and the classification is output together with the document.

The system administrator can access the document classifier directly by via path 70 to allow switching between the knowledge acquisition sub-modes and the document classification mode. Such switching may be used, for example, if a mis-directed document has been returned to the system administrator. The system administrator may then cause the document classifier to enter the learning sub-mode of knowledge acquisition mode, the system administrator inputting the correct classification for the document to the

classifier 30 together with the document to the feature extraction module 20, from which the features are extracted and passed to the classifier 30, so that the mis-directed document and associated correct classification are added to the knowledge base.

Similarly, at any point in operation of the document classification apparatus, the systems administrator can add additional training documents and/or rules by switching from document classification mode to knowledge acquisition mode.

The highlighted processes will now be explained in more detail.

Feature Selection

For document classification, there is a need to represent text documents in some format- and language- independent form, commonly termed a *feature* representation, before processing by a classifier. One of the most common forms of representation of features is that of singular word tokens. Specifically, the tokens are individual words that have been extracted from each document and transformed to their root form (e.g. root form of "selection" is "select"). Other "filtering" options based on sentence structure, such as recognizing only nouns while ignoring other word types such as prepositions and conjunctions, can also be used as will be apparent to those skilled in the art, in dependence upon requirements.

The keyword-based feature sets can be pre-defined manually or generated automatically from a set of pre-labeled documents.

The algorithm for automatic keyword selection accepts a list of pre-classified (i.e. training) text documents which are analyzed, processing one document at a time. Processing involves the extraction of all nouns (in root form) from each document and recording the number of occurrences of each of these prospective keywords within each category as well as within each document. Based on a certain set of selection rules, an overall rating of the "quality" of each word as a keyword is calculated and the list of keywords sorted by this value. The top N keywords with the highest rating are then selected as the "feature space" to be used for representing all documents (training or otherwise). The algorithm uses four different selection rules in ranking keywords which are combined to form a selection rating (f_{rating}). These are:

(a) Class Entropy

(b) Document Entropy

(c) Relative Document Count

(d) Document Inclusion Rate

a) Class Entropy (f_{CE}): this measures the distribution of a keyword's occurrence across the different categories. The more "polarized" the keyword's occurrence is towards a particular category, the more significant will the keyword be. This is because a keyword which occurs almost solely within one category and not at all in the others is much more likely to have some non-trivial association with the that category, as compared with a keyword which has a more even distribution across the categories.

The formula used to calculate class entropy for C different categories is:

$$f_{CE} = 1 - \sum_{i=1}^{i=C}\left(\left(Count(i) / \sum_{j=1}^{j=C} Count(j)\right) \times \log(Count(i)\right)$$

where:

$Count(x)$ = Total number of occurrences of keyword in category x

b) Document Entropy (f_{DE}): this measures the distribution of a keyword's occurrence across the different documents in a particular category. The criteria for a good keyword here is the opposite of that for Class Entropy. Here, the keyword which is much more evenly distributed across the documents in one category is a much better feature than one that has a more "polarized" distribution. This is because a keyword that occurs in more documents within a category is more likely to be one more commonly associated with documents of that category.

The formula used to calculate document entropy for D documents within 1 category, is:

$$f_{DE} = \sum_{i=1}^{i=D}\left(\left(Count(i) / \sum_{j=1}^{j=D} Count(j)\right) \times \log(Count(i)\right)$$

where:

$Count(x)$ = Total number of occurrences of keyword in document x

c) Relative Keyword Count (f_{RKC}): for a particular keyword, the top 2 document categories are defined as the 2 categories with the highest absolute count for that keyword. The

keyword-per-document ratio ($f_{Ratio\ i}$) for a category, i, is the total keyword count (C_i) for the category divided by the total number of documents (D_i) in that category. This relation can be expressed simply as:

$$f_{Ratio\ i} = C_i / D_i$$

Relative Keyword Count thus gives an indication of the difference between the keyword-per-document ratio of the 1st (f_{Ratio1}) and 2nd (f_{Ratio2}) categories. A keyword with a large difference between f_{Ratio1} and f_{Ratio2} is better than one with a small difference.

A measurement of f_{RKC} for C different categories is given by:

$$f_{RKC} = (f_{Ratio1} - f_{Ratio2}) / f_{Ratio1}$$

d) Document Inclusion Rate (f_{DIR}): f_{RDC} can be skewed by the high number of occurrences of a keyword in just one or two documents of a category. The use of f_{DIR} helps to "balance out" such situations by considering the number of documents in the top category in which the keyword occurs at least once.

A measurement of f_{DIR} for D_{1st} documents in the top category is given by:

$$f_{DIR} = O_{1st} / D_{1st}$$

where:

O_{1st} = number of documents in top category in which keyword occurs.

The overall ranking of each keyword is therefore simply derived by taking:

$$f_{Ranking} = f_{CE} \times f_{DE} \times f_{RDC} \times f_{DIR}$$

with:

$$0.0 <= f_{Ranking} <= 1.0$$

In this case, equal weightage has been given to each factor. Different coefficients could easily be added to each factor to give it a larger or smaller weightage.

The following example uses a small training set of two categories with 124 relevant documents each. The categories are business newspaper articles in the first category and non-business (e.g. sports) articles in the other. Consider a sampling of 40 keywords taken from the set of all keywords selected from the training sets. The total count of each keyword within each category as well as the number of documents (per category) in which it occurs, is as shown in Table 1. In Table 2, the "paths" (as shown by arrowed lines) of two exemplary keywords as they are ranked according to the four different factors, together with the final rating are shown with the combination of the four factors helping to provide a better overall view of the relative suitability of each keyword.

TABLE 1

Keyword	Total count		Unit document count	
	Cat 1	Cat 2	Cat 1	Cat 2
annual	33	2	22	2
authority	27	2	23	2
bank	92	6	28	4
capital	61	1	29	1
cent	560	32	87	21
champion	0	60	0	31
coach	0	26	0	19
company	191	12	65	8
corporate	48	4	26	3
cup	0	54	0	20
distribute	17	0	15	0
economy	87	1	28	1
event	2	34	1	25
exchange	34	2	23	2
fan	1	36	1	18
final	12	98	10	45
game	1	66	1	32
industry	137	6	59	6
invest	99	2	46	2
mainboard	26	0	22	0
market	175	4	64	3
match	6	65	3	28
minister	41	3	23	3
pe	499	43	91	21
play	12	181	11	55
potential	31	0	23	0
profit	78	5	28	3
property	96	3	37	3
rate	113	2	34	2
round	0	60	0	28
score	0	30	0	17
share	217	8	47	8
star	0	49	0	23
stock	104	0	40	0
technology	50	0	26	0
tournament	0	34	0	19
venture	53	1	20	1
victory	0	30	0	20
win	9	149	5	60
Woman	0	50	0	22

TABLE 2

	f_{CE}		f_{DE}		f_{RDC}		f_{DR}		f_{Rating}
champion	1.000	cent	0.872	champion	1.000	pe	0.734	market	0.331
coach	1.000	pe	0.872	coach	1.000	cent	0.702	cent	0.319
cup	1.000	company	0.815	cup	1.000	company	0.524	pe	0.261
distribute	1.000	market	0.802	distribute	1.000	market	0.516	industry	0.254
mainboard	1.000	win	0.802	mainboard	1.000	win	0.484	company	0.247
potential	1.000	industry	0.781	potential	1.000	industry	0.476	win	0.233
round	1.000	play	0.767	round	1.000	play	0.444	invest	0.231
score	1.000	invest	0.752	score	1.000	share	0.379	stock	0.223
star	1.000	share	0.740	star	1.000	invest	0.371	share	0.198
stock	1.000	final	0.733	stock	1.000	final	0.363	play	0.192
technology	1.000	stock	0.690	technology	1.000	stock	0.323	champion	0.170
tournament	1.000	champion	0.682	tournament	1.000	property	0.298	property	0.154
victory	1.000	game	0.680	victory	1.000	rate	0.274	game	0.152
woman	1.000	property	0.679	woman	1.000	game	0.258	rate	0.150
economy	0.910	event	0.652	economy	0.981	champion	0.250	round	0.137
game	0.888	match	0.649	game	0.977	capital	0.234	technology	0.129
capital	0.881	rate	0.647	capital	0.976	round	0.226	capital	0.128
rate	0.873	capital	0.636	venture	0.973	profit	0.226	economy	0.124
venture	0.867	authority	0.633	rate	0.967	match	0.226	potential	0.117
invest	0.860	potential	0.629	fan	0.964	economy	0.226	mainboard	0.111
market	0.846	bank	0.626	invest	0.964	bank	0.226	star	0.108
fan	0.821	mainboard	0.623	market	0.947	technology	0.210	final	0.107
property	0.804	corporate	0.620	property	0.946	corporate	0.210	woman	0.105
share	0.778	exchange	0.620	event	0.926	event	0.202	victory	0.097
industry	0.749	economy	0.617	exchange	0.926	star	0.185	tournament	0.091
cent	0.697	technology	0.615	annual	0.924	potential	0.185	coach	0.090
exchange	0.690	annual	0.610	industry	0.912	minister	0.185	cup	0.088
event	0.690	profit	0.610	authority	0.911	exchange	0.185	event	0.084
win	0.685	round	0.609	minister	0.905	authority	0.185	bank	0.084
annual	0.684	minister	0.602	share	0.905	woman	0.177	profit	0.083
company	0.676	victory	0.602	profit	0.900	mainboard	0.177	score	0.076
profit	0.672	tournament	0.595	bank	0.892	annual	0.177	venture	0.075
bank	0.668	woman	0.592	corporate	0.888	victory	0.161	match	0.074
play	0.664	coach	0.587	win	0.876	venture	0.161	exchange	0.074
minister	0.641	star	0.582	match	0.866	cup	0.161	corporate	0.070
authority	0.638	fan	0.564	company	0.854	tournament	0.153	authority	0.068
corporate	0.609	score	0.556	play	0.851	coach	0.153	annual	0.068
pe	0.600	distribute	0.554	final	0.800	fan	0.145	distribute	0.067
match	0.582	venture	0.551	cent	0.749	score	0.137	minister	0.065
final	0.503	cup	0.549	pe	0.679	distribute	0.121	fan	0.065

The algorithm allows for the specification of a minimum number K, of non-zero keyword counts which are expected to be found within each training document. The training documents are pre-processed by the method described above to determine the number of non-zero keyword counts in each document. Whenever a training document is found to have too few non-zero keyword counts, the next highest ranked keywords within the document are added to the set of N keywords initially selected, to bring the number of non-zero keyword counts for that document up to K. The total number of unique "bonus" keywords B, extracted from all training documents thus increases the dimension of the feature space to N+B.

Keyword Extraction

Once the keywords have been selected in the manner described above, keywords are extracted from a document and are formed into a feature vector, using the N+B set of keywords obtained during the selection process as the limited set of significant keywords that are to be searched within the document. This procedure is applied to both training documents to produce a set of respective training feature vectors and new documents to produce, respective feature vectors for yet-to-be categorized documents.

Based on the selected keyword features, the feature extraction algorithm parses the document to record the number of times a keyword w_i appear in the document (c_i). The keyword counts are then normalized such that the maximum score is 1 and the minimum score is 0. These scores are then provided as input to the classifier 30 as a normalized

keyword frequency count feature vector which encodes the statistical distribution of the keywords in the documents and thus provides a rough representation of the document content.

The feature extraction process using two sample articles is illustrated below. The first article, for Category 1 (business section) produces a positive word count for certain predominantly business-related keywords which are converted to relative frequency values as shown to form the input vector. The second article, for Category 2 (sports, music and life section) produces a positive word count for certain predominantly sports-related keywords which are likewise converted to relative frequency values to from the input vector.

Sample article for Category 1 (business section)

JUN 30 1997 Stationery maker Nippecraft in the red

MAINBOARD-LISTED specialist stationery maker Nippecraft has reported losses of $12 million for the year ended March, but said a company reorganisation would improve its bottomline this year.

The losses came on the back of a 4 per cent drop in turnover to $64 million and include exceptional and ordinary charges totalling close to $11 million, according to the company's unaudited results.

There will be no dividend payouts this year. Net tangible asset backing per ordinary share dropped to 0.69 cent, from 7.82 cents last year. The results were in line with Nippecraft's projections announced in February.

Managing director Bill Habergham attributed a large part of the loss to reorganisation of businesses in Britain, the United States, Australia and Malaysia.

"The exercise has now largely been completed and notwithstanding the tougher prospects ahead, we expect to reap the benefits of the reorganisation and restructure exercise in the current financial year," a company statement quoted him as saying.

Nippecraft said the exceptional charges, amounting to $6.8 million, included the writing down of stock by $5 million, operating losses and costs associated with the closure or restructuring of subsidiaries which cost close to $2 million.

The group managed to reduce inventory levels by a third to $18 million. Mr Habergham said this would benefit the group in the long term.

Keyword Table for Sample Article for Category 1

Keyword	Count	Relative Frequency
market	0	0.0
cent	**3**	**1.0**
pe	**2**	**0.7**
industry	0	0.0
company	**3**	**1.0**
invest	0	0.0
develop	0	0.0
stock	**1**	**0.3**
share	**1**	**0.3**
list	**1**	**0.3**
property	0	0.0
technology	0	0.0
capital	0	0.0
economy	0	0.0
sector	0	0.0
billion	0	0.0
potential	0	0.0
mainboard	**1**	**0.3**
project	**1**	**0.3**
play	0	0.0
win	0	0.0
champion	0	0.0
game	0	0.0
rate	0	0.0
round	0	0.0
star	0	0.0
final	0	0.0
woman	0	0.0
victory	0	0.0
tournament	0	0.0

Sample article for Category 2 (sports, music & life section)

JUL 2 1997 Love-fit Testud clinches famous win

LONDON -- One point short of a famous victory, Sandrine Testud rolled her eyes to the leaden skies then across the net to Monica Seles, shifting nervously from foot to foot.

Finally she turned to her Italian boyfriend Vittorio, huddled among the spectators overlooking the No. 3 court.

On his signal she served straight, deep and fast for her sixth ace and an astonishing 0-6, 6-4, 8-6 win over the Wimbledon second seed.

Nothing had looked less likely half-an-hour into Monday's third-round match.

On a court of low and uncertain bounce after the heavy rain which ravaged the opening week, Seles breezed through the first set in exactly 30 minutes.

"I was just trying to start playing," said the unseeded Testud.

"I was just so slow and nothing was going right."

The 25-year-old Frenchwoman, who lives and trains in Rome, finally won a game, holding serve in the second game of the second set after dropping the first two points.

She broke Seles in the next game. Hitting longer and with more power as she gained in confidence, Testud held service to win the second set in 42 minutes.

"I got a little bit tight, missed a couple of shots and the set was gone," Seles reflected.

Vittorio's contribution apparently goes further than his court-side advice.

How have you gotten so fit?, Testud was asked.

Keyword Table for Sample Article for Category 2

Keyword	Count	Relative Frequency
Market	0	0.0
Cent	0	0.0
Pe	0	0.0
Industry	0	0.0
Company	0	0.0
Invest	0	0.0
develop	0	0.0
stock	0	0.0
share	0	0.0
list	0	0.0
property	0	0.0
technology	0	0.0
capital	0	0.0
economy	0	0.0
sector	0	0.0
billion	0	0.0
Potential	0	0.0
Mainboard	0	0.0
Project	0	0.0
Play	0	0.0
Win	**4**	**1.0**
Champion	0	0.0
Game	**3**	**0.8**
rate	0	0.0
round	**1**	**0.2**
star	0	0.0
final	**2**	**0.5**
woman	0	0.0
victory	**1**	**0.2**
tournament	0	0.0

<u>The Classifier: Adaptive Resonance Associative Map (ARAM)</u>

ARAM is a family of neural network models that performs incremental supervised learning of recognition categories (pattern classes) and multidimensional maps of both binary and analog patterns. An ARAM system is shown in Figure 4 and can be visualized as two overlapping Adaptive Resonance Theory (ART) [1,2,3] modules consisting of two input fields F_1^a (300) and F_1^b (310) with an F_2 category field (320). For classification problems, the F_1^a field (300) serves as the input field containing the input activity vector and the F_1^b field (310) servers as the output field containing the output class vector. The F2 field (320) contains the activities of categories that are used to encode the patterns. During learning, given an input pattern presented at the F_1^a input layer and an output pattern presented at the F_1^b output field, a F_2 category node is selected to encode the pattern pair.

When performing classification tasks, ARAM formulates recognition categories of input patterns, and associates each category with its respective prediction. The knowledge that ARAM discovers during learning is compatible with IF-THEN rule-based representation. Specifically, each node in the F_2 field (320) represents a recognition category associating the F_1^a input patterns with the F_1^b output vectors. Learned weight vectors, one for each F_2 node, constitute a set of rules that link antecedents to consequents. At any point during the incremental learning process, the system architecture can be translated into a compact set of rules. Similarly, domain knowledge in the form of IF-THEN rules can be inserted into ARAM architecture.

The ART modules used in ARAM can be ART 1 [1], which categorizes binary patterns, or analog ART modules such as ART 2-A [2], and fuzzy ART [3], which categorize both binary and analog patterns. The fuzzy ARAM model, that is composed of two overlapping fuzzy ART modules is described below.

Knowledge Acquisition Mode

Learning Sub-Mode

In the learning sub-mode of knowledge acquisition mode, ARAM learns a set of recognition categories or rules by training from pre-labeled document sets. During learning, the keyword frequency vectors, each representing a document, are presented to ARAM as input **A** one at a time together with the associated class label input **B**.

Given an input keyword vector **A**, ARAM first searches for a F_2 recognition category encoding a keyword template vector that is closest to the input vector according to some similarity measure. It then checks if the associated F_2 output prediction of the selected category matches with the output class label **B**. If so, under fast learning, the keyword templates of the F_2 recognition category is modified to contain the intersection of the original keyword templates and the input keyword vector. Otherwise, the recognition category is reset and the system repeats to select another category until a match is found.

Given a set of the documents with a specific class label, the system learns to pick up combinations of keywords that consistently appear in the documents and derive rules that associate combinations of the relevant keywords to the target output class of the

documents. ARAM learning is stable in the sense that weight values do not oscillate, as they can only decrease but not increase. As new cases are incorporated by adjusting the weight templates of the chosen category nodes, learning does not wash away previously learned knowledge. This allows incremental learning in the sense that learning of new cases does not require relearning of old data.

The detailed algorithm for incremental learning is given below:

Input vectors: The F_1^a and F_1^b input vectors are normalized by *complement coding* that preserves amplitude information. Complement coding represents both the on-response and the off-response to an input vector **a**. The complement coded F_1^a input vector **A** is a 2M-dimensional vector

$$\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) = (a_1,\ldots,a_M,a_1^c,\ldots,a_M^c)$$

Where a_i represents the normalized frequency score of keyword w_i, and $a_i^c = 1 - a_i$.

Similarly, the complement coded F_1^b input vector **B** is a 2N-dimensional vector

$$\mathbf{B} = (\mathbf{b}, \mathbf{b}^c) = (b_1,\ldots,b_N,b_1^c,\ldots,b_N^c)$$

where b_i represents the presence ($b_k=1$) or absence ($b_k=0$) of a class label c_k, and $b_k^c = 1 - b_k$.

<u>Activity vectors:</u> Let $\mathbf{x}^a$ and $\mathbf{x}^b$ denote the F_1^a and F_1^b activity vectors respectively. Let $\mathbf{y}$ denote the F_2 activity vector.

<u>Weight vectors:</u> Each F_2 category node j is associated with two adaptive weight templates $\mathbf{w}_j^a$ and $\mathbf{w}_j^b$. Initially, all category nodes are uncommitted and all weights equal ones. After a category node is selected for encoding, it becomes committed.

Fuzzy ARAM dynamics are determined by the choice parameters $\alpha_a > 0$ and $\alpha_b > 0$; the learning rates β_a in [0,1] and β_b in [0,1]; the vigilance parameters ρ_a in [0,1] and ρ_b in [0,1]; and a contribution parameter γ in [0,1].

<u>Bottom up propagation:</u> Given the F_1^a input vectors $\mathbf{A}$, for each F_2 node j, the F_1^a to F_2 input T_j is defined by :

$$T_j = |\mathbf{A} \wedge \mathbf{w}_j^a| \,/\, (\alpha_a + |\mathbf{w}_j^a|)$$

where the fuzzy AND operation $\wedge$ is defined by $(\mathbf{p} \wedge \mathbf{q})_i = \min(p_i, q_i)$, and where the norm $|.|$ is defined by $|\mathbf{p}| = \Sigma_i\, p_i$ for vectors $\mathbf{p}$ and $\mathbf{q}$.

<u>Category choice:</u> Using a choice rule, at most one F_2 node can become active. The choice is indexed at J where $T_J = \max\{T_j\text{: for all } F_2 \text{ node } j\}$.

When a category choice is made at node J, $y_J = 1$; and $y_j = 0$ for all j not equal to J.

<u>Resonance or reset:</u> Resonance occurs if the *match functions*, m_J^a and m_J^b, meet the vigilance criteria in their respective modules:

$$m_J^a = |\mathbf{A} \wedge \mathbf{w}_J^a| \,/\, |\mathbf{A}| >= \rho_a \text{ and } m_J^b = |\mathbf{B} \wedge \mathbf{w}_J^b| \,/\, |\mathbf{B}| >= \rho_b.$$

Learning then ensues, as defined below. If any of the vigilance constraints is violated, mismatch reset occurs in which the value of the choice function T_J is set to 0 for the duration of the input presentation. The search process repeats to select another new index J until resonance is achieved.

<u>Learning :</u> Once the search ends, the weight vectors $\mathbf{w}_J^a$ and $\mathbf{w}_J^b$ are updated according to the equations

$$\mathbf{w}_J^{a\,(new)} = (1-\beta_a)\, \mathbf{w}_J^{a\,(old)} + \beta_a\, (\mathbf{A} \wedge \mathbf{w}_J^{a\,(old)})$$

and

$$\mathbf{w}_J^{b\,(new)} = (1-\beta_b)\, \mathbf{w}_J^{b\,(old)} + \beta_b\, (\mathbf{B} \wedge \mathbf{w}_J^{b\,(old)})$$

respectively. For efficient coding of noisy input sets, it is useful to set $\beta_a = \beta_b = 1$ when J is an uncommitted node, and then take $\beta_a < 1$ and $\beta_b < 1$ after the category node is committed. *Fast learning* corresponds to setting $\beta_a = \beta_b = 1$ for committed nodes.

Match tracking: At the start of each input presentation, the vigilance parameter ρ_a equals a baseline vigilance $\underline{\rho}_a$. If a reset occurs in the category field F_2, ρ_a is increased until it is slightly larger than the match function m_J^a. The search process then selects another F_2 node J under the revised vigilance criterion.

Rule Insertion Sub-Mode

Through the rule insertion process, user-defined knowledge in the form of rules is inserted into the ARAM network (knowledge base). A rule is typically in the IF-THEN format, consisting of a set of keyword features as the antecedents and a classification as the consequent. Due to the compatibility of ARAM architecture and rules, domain knowledge in the form of IF-THEN rules can be readily inserted into an ARAM network.

For example, given a rule such as

"Stock", "Share", "Price" -> Business,

the rule insertion algorithm creates a keyword frequent vector in which the frequency score of *"Stock"*, *"Share"* and *"price"* are 1s and all others zeros; and an output class vector in which the score of *"Business"* is 1 and all others zeros. Given the keyword frequency vector in the F_1^a field, and the class vector in the F_1^b field, ARAM first searches for a recognition category that encodes the exact same set of keywords. If such a recognition category exists and its predicted class is *"Business"*, no action is required as the rule already exists. If the predicted class is not "Business", a contradiction occurs and it is flagged to the users. If such a recognition category does not exist, a recognition

category is created to encode a keyword template consisting of *"Stock"*, *"Share"*, and *"Price"* and a classification of *"Business"*.

The detailed rule insertion algorithm is as follows:

Rule insertion proceeds in two phases. The first phase translates each rule into a 2M-dimensional vectors **A** and a 2N-dimensional vectors **B**, where *M* is the total number of document features and N is the number of classes.

In the most general case, given a rule of the following format,

IF $x_1, x_2, \ldots, x_m$, not(x'_1), not(x'_2), ..., not(x'_m)

THEN $y_1, y_2, \ldots, y_n$, not(y'_1), not(y'_2), ..., not(y'_m)

where $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ are positive attributes, and $x'_1, x'_2, \ldots, x'_m$ and $y'_1, y'_2, \ldots, y'_n$ preceded by the logical NOT operator are negative attributes, the algorithm derives the pair of vectors

$$\mathbf{A} = (\mathbf{a}, \mathbf{a}^c) \text{ and } \mathbf{B} = (\mathbf{b}, \mathbf{b}^c)$$

such that for each index j = 1, ...,M,

$(a_j, a_j^c) =$

(1,0) if $w_j = x_i$ for some i in {1, ...,m}

(0,1) if $w_j = x'_i$ for some i in {1, ...,m'}

(0,0) otherwise

and

(b_k, b^c_k) =

(1,0) if $c_k = y_i$ for some i in {1,...,n}

(0,1) if $c_k = y'_i$ for some i in {1,...,n'}

(0,0) otherwise

where w_j is the j^{th} keyword feature and c_k is the k^{th} class label.

The vector pairs derived from the rules are then used as training patterns to initialize an ARAM network. Given a pair of vectors **A** and **B** derived from a rule, their respective recognition categories are associated through the map field.

During network initialization, the vigilance parameters ρ_a and ρ_b are each set to 1 to ensure that only identical attribute vectors are grouped into one recognition category. Contradictory symbolic rules are detected during rule insertion when identical input attribute vectors are associated with distinct output attribute vectors. The detection is achieved through a perfect mismatch phenomenon, in which the system tries to raise ρ_a above 1 in response to a mismatch in F_1^a.

Document classification

Given an input document, a feature extraction module parses the text to derive a normalized keyword frequency vector (as described above). The complement coded input vector **A** is then presented to the F_1^a field.

Given an input keyword vector **A**, ARAM first searches for a F_2 recognition category encoding a keyword template vector that is closest to the input vector according to the choice function. The associated F_2 output prediction of the selected category is then used as the output class label.

Choice Rule: In ARAM systems with category choice, only the F_2 node J that receives maximal F_1^a to F_2 input T_j predicts output. Specifically:

$y_j =$ 1 if $j = J$ where $T_J > T_k$ for all k not equal to J;

0 otherwise

The F_1^b activity vector $\mathbf{x}^b$ is given by $\mathbf{x}^b = \Sigma_j \mathbf{w}_j^b y_j = w_J^b$ and the output vector $\mathbf{B}_2 = (b_1, b_2, \ldots, b_N)$ is then read directly from the F_1^b field. The output class is predicted to be K if $b_K > b_k$ for all k not equal to K and the confidence value is given by b_K.

Confidence Value

Given training examples and rules of a single class output and with fast learning, ARAM associates input features to a binary class prediction. In other words, only one output class b_K equals one and $b_k = 0$ for all k not equal to K. To derive a real value prediction score between 0 and 1, a few strategies are possible, of which two are described below:

a) Distributed category prediction

Using distributed category prediction, more than one F_2 nodes can become active. The F_2 output vector y represents a less extreme contrast enhancement of the F_1^a to F_2 input **T**, in the sense that the higher T_j's are amplified and smaller T_j's are further reduced. Two algorithms that approximate contrast enhancement are given below.

Power Rule: The power rule raises the input T_j to the j^{th} F^2 node to a power p and normalizes the total activity:

$$y_j = (T_j)^p / \ \Sigma_k (T_k)^p.$$

The power rule converges toward the choice rule as p becomes large.

K-max Rule: In the spirit of the K Nearest Neighbor (KNN) system, the K-max rule picks the set of K F_2 nodes with the largest input T_j for prediction. The F_2 activity values y_j are then:

$$y_j = T_j / \Sigma_{k \text{ in } \pi} T_k \quad \text{if } j \text{ in } \pi$$
$$0 \qquad \text{otherwise,}$$

where π is the set of K category nodes with the largest T_j values. The K-max rule with K=N is equivalent to the power rule with p=1.

After the F_2 activity vector **y** is contrast enhanced by the power rule or the K-max rule, the output activity vector $\mathbf{x}^b$ in the F_1^b field computed by

$$\mathbf{x}^b = \Sigma_j \mathbf{w}_j^b \; y_j$$

The output vector $\mathbf{B}_2 = (b_1, b_2, \ldots, b_N)$ is then read directly from $\mathbf{x}^b$. The output class is predicted to be K if $b_K > b_k$ for all k not equal to K and the confidence value is given by b_K.

b) Voting strategy

Using the voting strategy technique, multiple ARAM systems are inserted with different sets of rules and/or trained using different sets of input patterns or different orderings of the input patterns. When performing classification, each ARAM votes for its predicted class. The voting scores normalized by the number of ARAM provide a prediction score between 0 and 1 for each output class.

$$s_j = v_j / \Sigma_k v_k$$

where v_j is the number of votes given to and s_j is the normalized prediction score for the output class j. The output class with the highest prediction score is the selected predicted class and its prediction score is the confidence value.

Switching between modes

The system administrator can switch between the classification mode and the knowledge acquisition sub-modes by sending a message together with the appropriate data to the document classifier. The message can be either LEARN, INSERT, or CLASSIFY. Depending on the message received, the document classifier adjusts the input baseline vigilance parameter $\underline{\rho}_a$ and the output vigilance parameter ρ_b of the ARAM classifier accordingly and carries out the appropriate sequence of actions.

With a LEARN message, the document classifier receives a document text together with a classification label. First, the feature extraction module derives the normalized keyword frequency vector from the document. The keyword vector is presented as the input vector to the F_1^a field and the classification vector (based on the classification label) is presented to the F_1^b field of the ARAM classifier. The ARAM classifier is then run with $\underline{\rho}_a < 1$ (typically 0, to maximize generalization) and $\rho_b = 1$.

With an INSERT message, the document classifier receives an IF-THEN rule. First, the rule insertion module converts the given rule into a pair of input and output vectors, presents the

input vector to the F_1^a field and the output vector to the F_1^b field. The ARAM classifier is then run with both the input and output vigilance parameters set to 1s.

With a CLASSIFY message, the document classifier receives a document text. First, the feature extraction module derives the normalized keyword frequency vector from the document and presents it as the input vector to the F_1^a field. The ARAM classifier is then run with both ρ_a and ρ_b equal to zeroes to ensure a prediction is always made. The predicted classification label is then read from the F_1^b field and returned to the user.

Figures 5, 6, and 7 summarize the parameter setting and the relevant functional blocks of the document classification system in the learning, rule insertion, and document classification modes respectively.

The embodiment described is not to be construed as limitative. For example, elthough the classifier module has been shown implemented using an ARAM structure, this may be implemented using any other structure which allows incremental learning and rule insertion, such as an ARTMAP structure.

References:

[1] G. A. Carpenter & S. Grossberg (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, **37**, 54-115.

[2] G. A. Carpenter, S. Grossberg & D. B. Rosen (1991a). ART 2-A: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 493-504.

[3] G. A. Carpenter, S. Grossberg & D. B. Rosen (1991b). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759-771.

[4] C. A. Carpenter & A.-H. Tan (1995). Rule Extraction: From Neural Architecture to Symbolic Representation'. *Connection Science*, 7(1), 3-27.

[5] A.-H. Tan (1995). Adaptive Resonance Associative Map. *Neural Networks*, **8**(3), 437-446.

[6] A.-H. Tan, ``Cascade ARTMAP: Integrating Neural Computation and Symbolic Knowledge Processing''. IEEE Transactions on Neural Networks, Vol. 8, No. 2 (March 1997) 237-250.

CLAIMS

1. Document classification apparatus comprising feature extraction means for extracting a plurality of features from a document and a classifier operable on the extracted features to process the document in a knowledge acquisition mode in which the association of a classification with the document is added incrementally to a knowledge base or in a document classification mode in which the classifier, using the knowledge base, determines a predicted classification for the document, the classifier being switchable between the modes under user control .

2. Apparatus as claimed in claim 1 wherein the classifier comprises a supervised adaptive resonance theory (ART) system.

3. Apparatus as claimed in claim 2 wherein the system comprises an ARTMAP system.

4. Apparatus as claimed in claim 2 wherein the system comprises an ARAM system.

5. Apparatus as claimed in any one of the preceding claims further comprising a router arranged to route the document to one of a plurality of destinations in dependence upon the classification.

6. Apparatus as claimed in any one of the preceding claims wherein the classification has associated therewith a confidence value.

7. Apparatus as claimed in claim 6 as dependent on claim 5 wherein the confidence value is comparable to a threshold, the router being arranged to make an automatic routing or manual routing decision in dependence upon the comparison.

8. Apparatus as claimed in claim 7 wherein the threshold is adjustable.

9. Apparatus as claimed in claim 7 or claim 8 wherein a said destination is a system administrator, responsible for manual routing.

10. Apparatus as claimed in any one of the preceding claims wherein the features are formed into a feature vector for input to the classifer.

11. Apparatus as claimed in any one of the preceding claims wherein the features comprise classification-associated words or phrases which may appear in the document.

12. Apparatus as claimed in any one of the preceding claims wherein the extracting means is arranged to provide a measure of the frequency of occurrence of the features in the document.

13. Apparatus as claimed in claim 5 wherein the destinations include a system administrator to which the other destinations are connected, mis-routed documents

being sendable by the other destinations to the system administrator for manual routing.

14. Apparatus as claimed in claim 13 wherein the system administrator is connected to the feature extraction means and classifier, the arrangement being such that a said mis-directed document, in association with an actual classification supplied by the system administrator, is processed in knowledge acquisition mode to add the association of the actual classification with the mis-directed document to the knowledge base.

15. Apparatus as claimed in any one of the preceding claims wherein the apparatus is operable to perform rule insertion in the knowledge acquisition mode in which a plurality of features are input by a user to the classifier together with a classification with which the features are associated.

16. Apparatus as claimed in any one of the preceding claims wherein the apparatus is operable in knowledge acquisition mode to process a plurality of training documents with associated classifications as a batch.

17. Document classification apparatus comprising:

feature extraction means for extracting a plurality of features from a document,

a classifier operable, using a knowledge base, to determine from the features a predicted classification for the document, the classification having a confidence value associated therewith; and

a router arranged to compare the confidence value to a threshold and make a decision to route the document automatically to one of a plurality of destinations or to a destination for manual routing in dependence upon the comparison.

17. Apparatus as claimed in claim 13 wherein the threshold is adjustable.

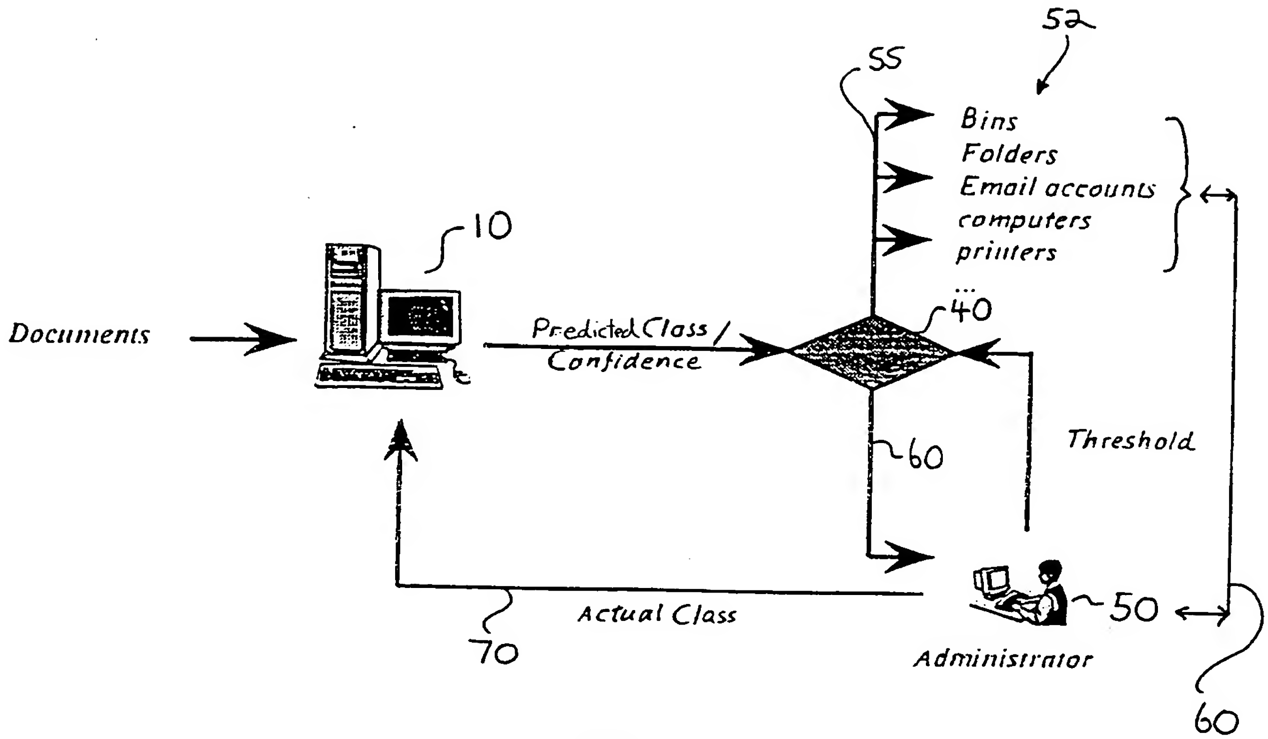

52
55
Bins
Folders
Email accounts
computers
printers
...
10
40
Documents
Predicted Class /
Confidence
Threshold
60
Actual Class
70
50
Administrator
60

FIG. 1

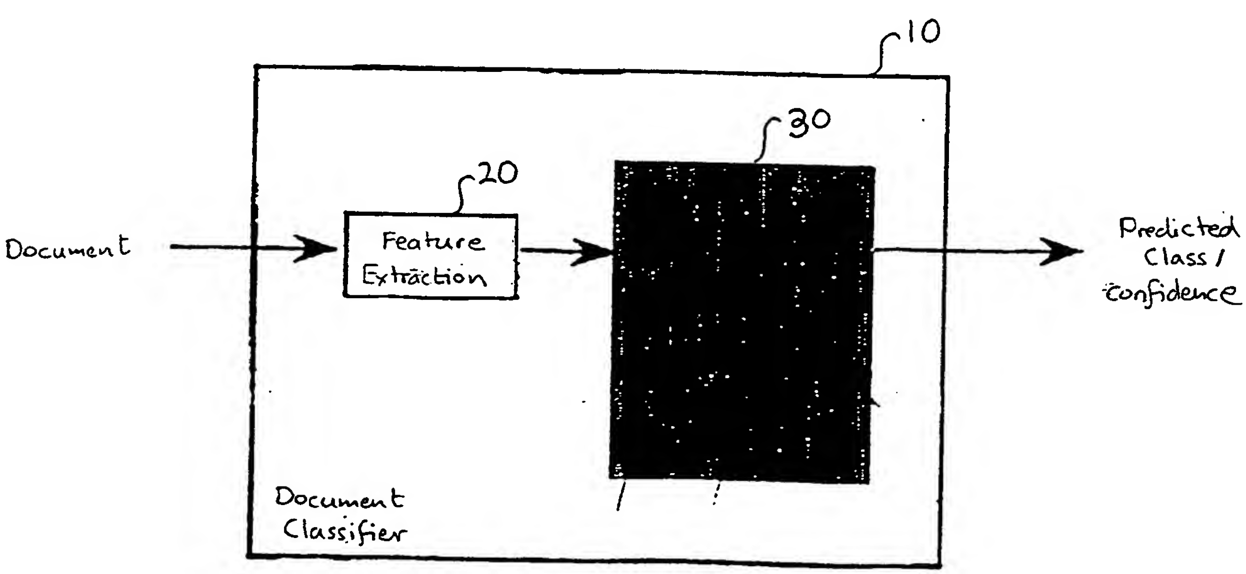
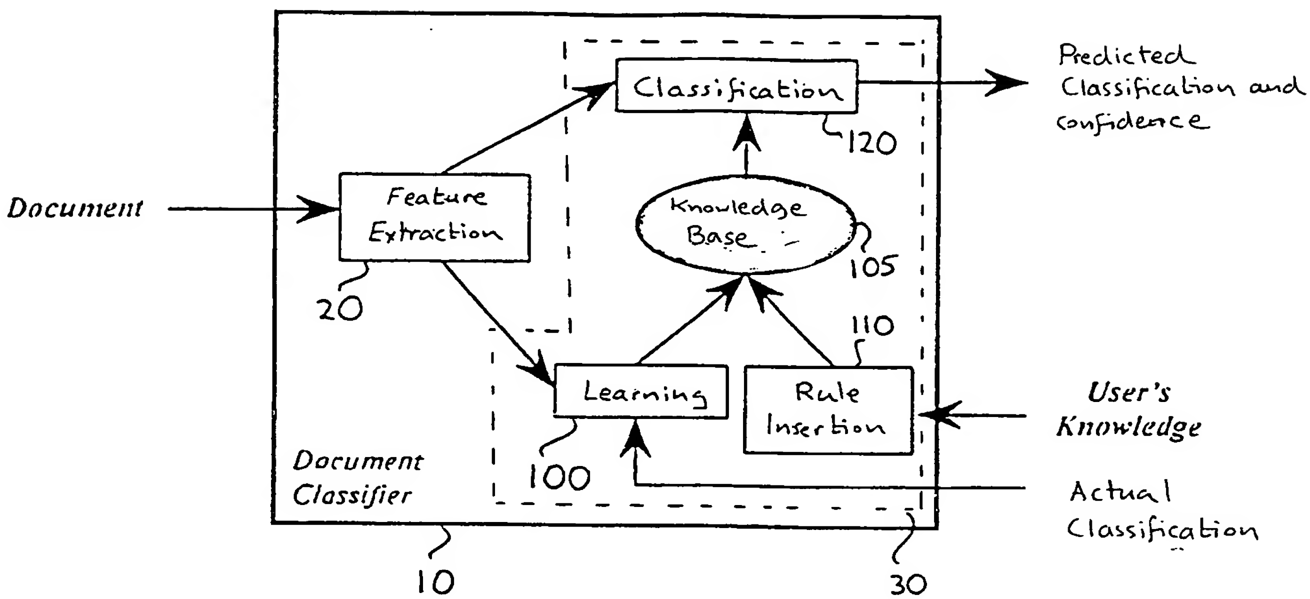

10
30
20
Document
Feature
Extraction
Predicted
Class /
confidence
Document
Classifier

FIG. 2



Document
Feature Extraction
20
Classification
120
Predicted Classification and confidence
Knowledge Base
105
110
Learning
Rule Insertion
100
User's Knowledge
Actual Classification
Document Classifier
10
30

FIG. 3

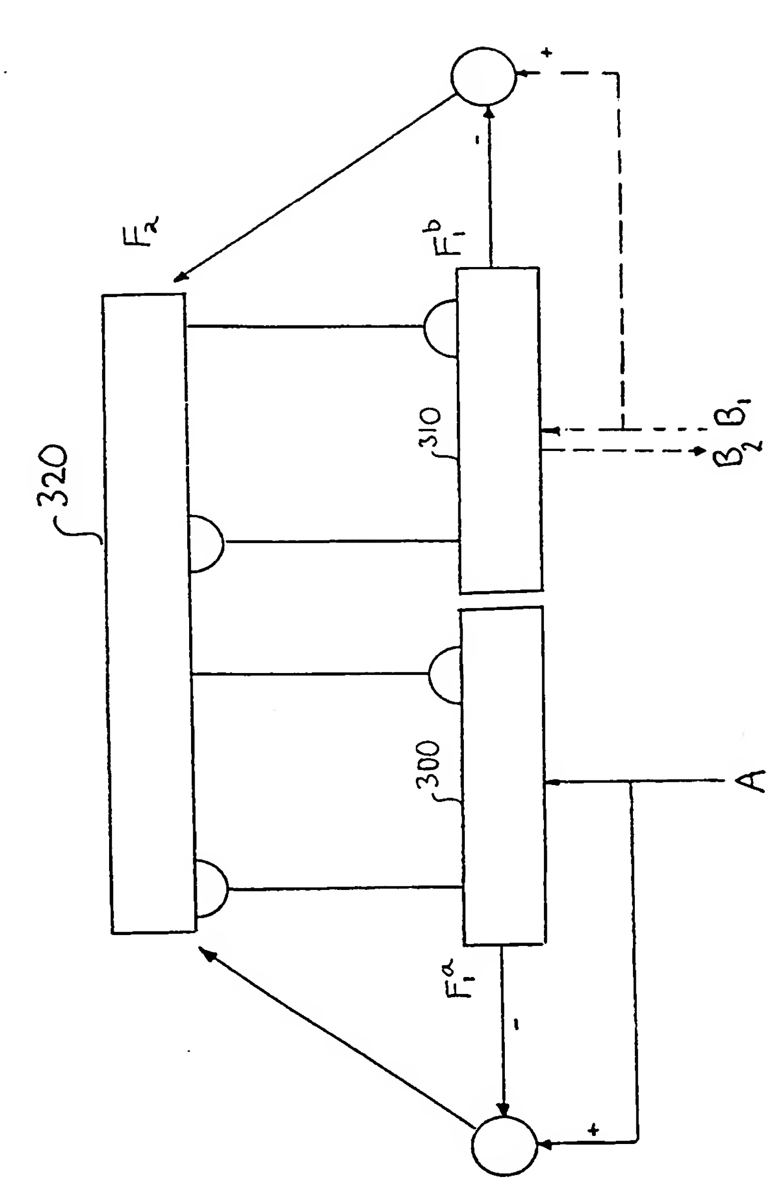


FIG. 4

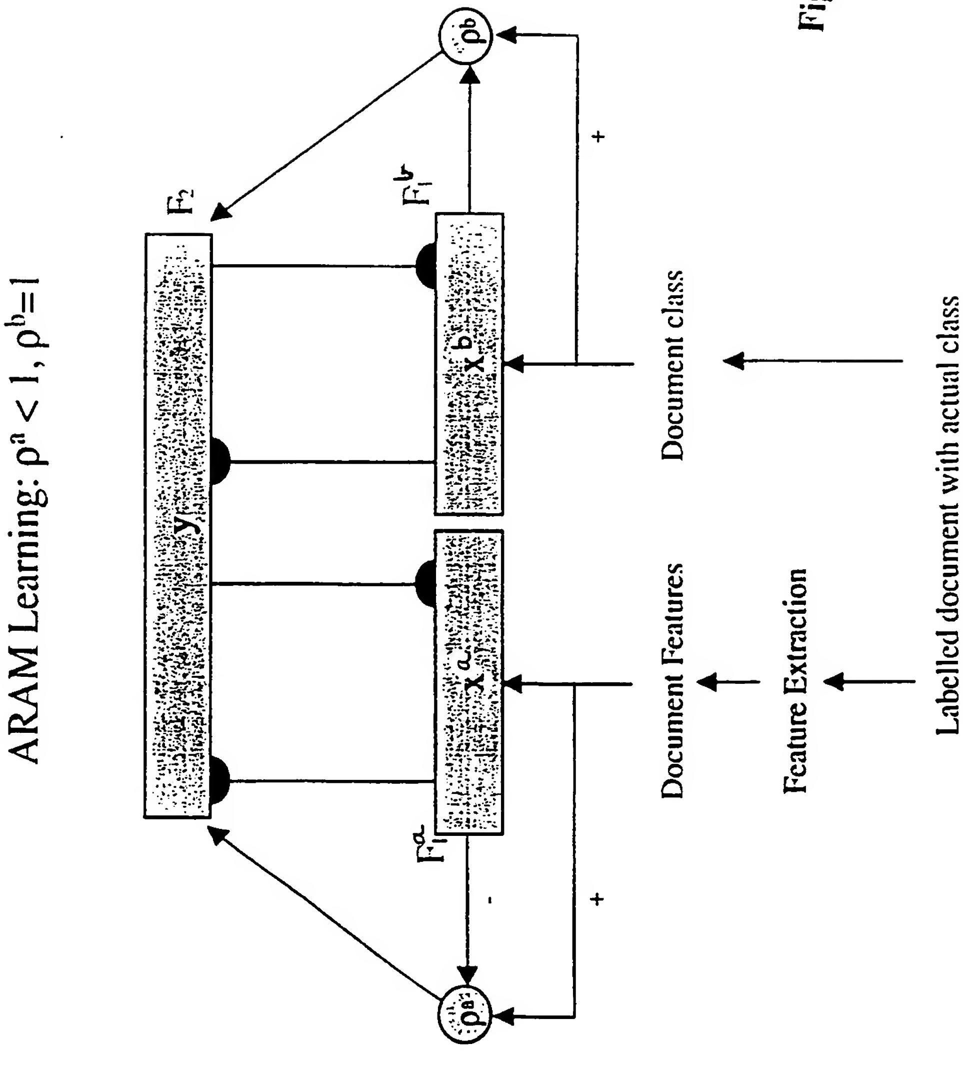

ARAM Learning: $\rho^a < 1$, $\rho^b = 1$
F_2
y
F_1^a
x^a
F_1^b
x^b
ρ^a
ρ^b
-
+
+
Document Features
Document class
Feature Extraction
Labelled document with actual class

Fig. 5

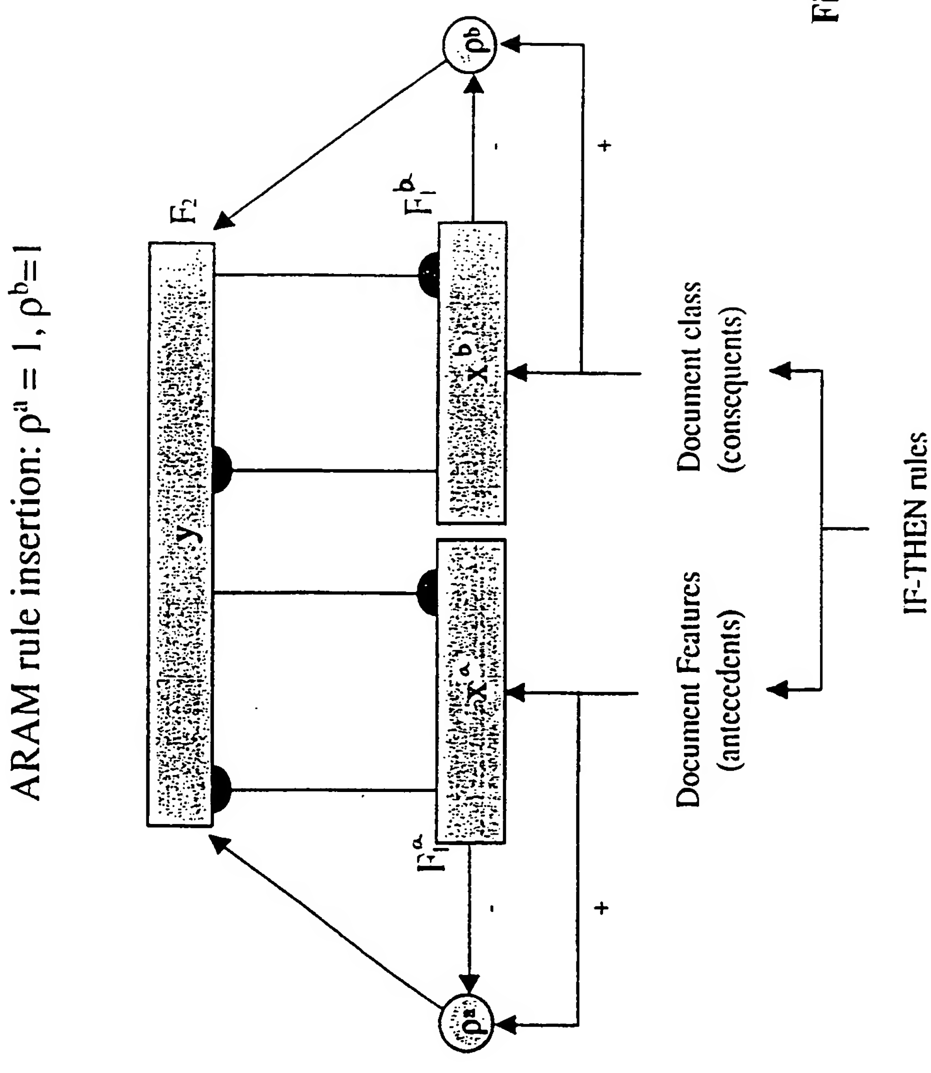

ARAM rule insertion: $\rho^a = 1$, $\rho^b=1$
F_2
y
F_1^a
F_1^b
x^a
x^b
ρ^a
ρ^b
-
+
-
+
Document Features (antecedents)
Document class (consequents)
IF-THEN rules

Fig. 6

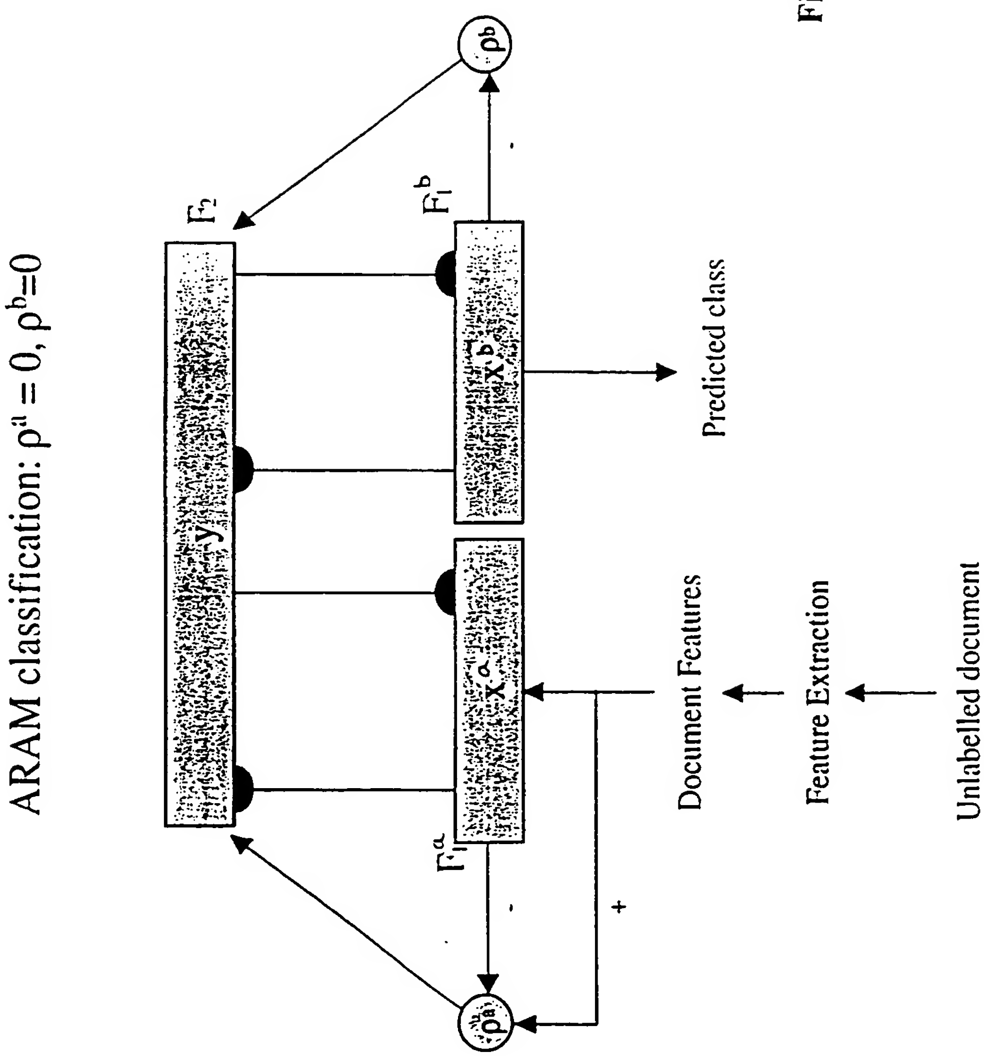

ARAM classification: $\rho^a = 0, \rho^b = 0$
F_2
y
F_1^a
x^a
F_1^b
x^b
ρ^a
ρ^b
+
-
-
Predicted class
Document Features
Feature Extraction
Unlabelled document

Fig. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/SG 99/00089

CLASSIFICATION OF SUBJECT MATTER
IPC[7]: G 06 F 15/80; G 06 K 9/66
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols) IPC[7]: G 06 F; G 06 K; H 04 N
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPODOC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 97/04400 A1 (THE COMMONWEALTH OF AUSTRALIA) 6 February 1997 (06.02.97) claims 1,6; fig.1-3.	1,2
A	US 5566273 A (HUANG et al.) 15 October 1996 (15.10.96) claims 1,2; fig.1-3.	1-3
A	US 5794236 A (MEHRLE) 11 August 1998 (11.08.98) abstract; fig.1-6.	1,17
A	GB 2278705 A (SPENCER) 7 December 1994 (07.12.94) abstract; fig.1.	1,17
A	US 5091964 A (SHIMOMURA) 25 February 1992 (25.02.92) claims 1,6; fig.1-3.	1,17
A	JP 11085796 A (CANON) 30 March 1990 (30.03.90) abstract. In: Patent Abstracts of Japan [CD-ROM].	1,10,17

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
„A" document defining the general state of the art which is not considered to be of particular relevance
„E" earlier application or patent but published on or after the international filing date
„L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
„O" document referring to an oral disclosure, use, exhibition or other means
„P" document published prior to the international filing date but later than the priority date claimed

„T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
„X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
„Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
„&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 July 2000 (11.07.2000)	2 August 2000 (02.08.2000)
Name and mailing adress of the ISA/AT Austrian Patent Office Kohlmarkt 8-10; A-1014 Vienna Facsimile No. 1/53424/535	Authorized officer Mihatsek Telephone No. 1/53424/329

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 11039313 A (NIPPON) 12 February 1999 (12.02.99) abstract; In: Patent Abstracts of Japan [CD-ROM].	1,17
A	JP 11085797 A (CANON) 30 March 1999 (30.03.99) abstract; In: Patent Abstracts of Japan [CD-ROM].	1,10,17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/SG 99/00089

Patent document cited in search report			Publication date	Patent family member(s)			Publication date
WO	A1	9704400	06-02-1997	AU	A0	4315/95	17-06-1995
				AU	A1	65094/96	16-02-1997
US	A	5566273	15-10-1996	FR	A1	2714748	07-07-1995
				FR	A1	2728365	21-06-1996
				FR	B1	2714748	23-08-1996
				FR	B1	2728365	23-05-1997
				JP	A2	7225748	22-08-1995
US	A	5794236	11-08-1998	AU	A1	30701/97	07-01-1998
				AU	B2	713225	25-11-1999
				CA	AA	2256408	18-12-1997
				EP	A1	970428	12-01-2000
				EP	A4	970428	14-06-2000
				WO	A1	9748057	18-12-1997
GB	A1	2278705	07-12-1994	GB	A0	9311256	21-07-1993
US	A	5091964	25-02-1992	GB	A0	9107066	22-05-1991
				GB	A1	2244886	11-12-1991
				GB	B2	2244886	10-08-1994
				JP	A2	3290774	20-12-1991
JP	A2	11085796	30-03-1999			none	
JP	A2	11039313	12-02-1999			none	
JP	A2	11085797	30-03-1999			none	